Surveillance Video Parsing with Single Frame Supervision

Si Liu[1], Changhu Wang[2], Ruihe Qian[1], Han Yu[1], Renda Bao[1], Yao Sun[1*]
[1]State Key Laboratory of Information Security, Institute of Information Engineering,
Chinese Academy of Sciences,Beijing 100093,China
[2]Toutiao AI Lab
{liusi, qianruihe, yuhan, sunyao}@iie.ac.cn, wangchanghu@toutiao.com

Abstract

Surveillance video parsing, which segments the video frames into several labels, e.g., face, pants, left-leg, has wide applications [41, 8]. However, pixel-wisely annotating all frames is tedious and inefficient. In this paper, we develop a Single frame Video Parsing (SVP) method which requires only one labeled frame per video in training stage. To parse one particular frame, the video segment preceding the frame is jointly considered. SVP ***(i)*** *roughly parses the frames within the video segment,* ***(ii)*** *estimates the optical flow between frames and* ***(iii)*** *fuses the rough parsing results warped by optical flow to produce the refined parsing result. The three components of SVP, namely frame parsing, optical flow estimation and temporal fusion are integrated in an end-to-end manner. Experimental results on two surveillance video datasets show the superiority of SVP over state-of-the-arts. The collected video parsing datasets can be downloaded via* `http://liusi-group.com/projects/SVP` *for the further studies.*

1. Introduction

In recent years, human parsing [16] is receiving increasing owning to its wide applications, such as person re-identification [41] and person attribute prediction [19, 38]. Most existing human parsing methods [15, 16, 37] target at segmenting the human-centric images in the fashion blogs. Different from fashion images, parsing surveillance videos is much more challenging due to the lack of labeled data. It is very tedious and time-consuming to annotate all the frames of a video, for a surveillance video usually contains tens of thousands of frames per second.

In this paper, we target at an important, practically applicable yet rarely studied problem: *how to leverage the very limited labels to obtain a robust surveillance video parsor*? More specifically, we mainly consider an extreme situation, i.e., only one frame in each training video is annotated. Note that labeling is unnecessary in testing phase. As shown in Figure 1, the labeled frame per training video (red bounding box) is fed into the proposed Single frame supervised Video Parsing (**SVP**) network. Insufficient labeled data always lead to over-fitting, especially in the deep learning based method. The rich temporal context among video frames can partially solve this problem. By building the *dense correspondences*, i.e., optical flow, among video frames, the single labeled frame can be viewed as the seed to indirectly expand (propagate) to the whole video. Most state-of-the-art optical flow estimation methods, such as EpicFlow [29] etc, suffer from relatively slow speed. Because the video parsing task requires extensive online optical flow computation, a real-time, accurate optical flow estimation is essential. Thus, it is a challenging but essential problem to build an end-to-end, efficient video parsing framework by only utilizing the limited (e.g, only one) labeled images and large amount of unlabeled images with online estimated dense correspondences among them.

To tackle these challenges, we propose the SVP network. As shown in Figure 1, to parse a test frame I_t, a parsing

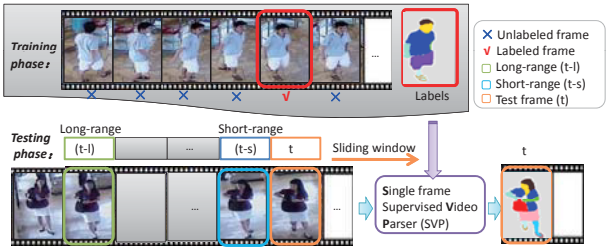

Figure 1. During training, only a single frame per video (red check mark) is labeled, while others (blue x mark) are unlabeled. A SVP network is learned from the extremely sparsely labeled videos. During testing, a parsing window is slided along the video. The parsing result of testing frame I_t (orange box) is determined by itself, the long-range frame I_{t-l} (green box) and the short-range frame frame I_{t-s} (blue box). For better viewing of all figures in this paper, please see original zoomed-in color pdf file.

*corresponding author

window which contains I_t and several frames preceding it $\{I_{t-k}, k = 0, ..., l\}$, is slided along the video. Considering the computation burden and cross-frame redundancies, a triplet $\{I_{t-l}, I_{t-s}, I_t\}$ is selected to represent the sliding window. The *long-range frame* I_{t-l} lies l frames ahead of I_t while *short-range frame* I_{t-s} lies s frames ahead of I_t. Usually, $l > s$. They complement each other in that the short-range optical flows are more accurate, while the long-range frames bring more diversities. The triplet is fed into SVP to collaboratively produce the parsing result.

SVP contains three sub-networks. The image parsing sub-network parses the three frames respectively, while the optical flow estimation sub-network builds the cross-frame pixel-wise correspondences. In order to decrease the interference of imperfect optical flow, a pixel-wise confidence map is calculated based on the appearance differences between one image and its counterpart wrapped from the other image. Based on the mined correspondences and their confidences, the temporal fusion sub-network fuses the parsing results of the each frame, and then outputs the final parsing result. Extensive experiments in the newly collected indoor and outdoor datasets show the superior performance of SVP than the state-of-the-arts.

The contributions of this work are summarized as follows. **(i)** To the best of our knowledge, it is the first attempt to segment the human parts in the surveillance video by labeling single frame per training video. It has extensive application prospect. **(ii)** The proposed SVP framework is end-to-end and thus very applicable for real usage. Moreover, the feature learning, pixelwise classification, correspondence mining and the temporal fusion are updated in a unified optimization process and collaboratively contribute to the parsing results.

2. Related Work

Image, video and part semantic segmentation: Long *et al.* [22] build a FCN that take input of arbitrary size and produce correspondingly-sized output. Chen *et al.* [4] introduce atrous convolution in dense prediction tasks to effectively enlarge the field of view of filters to incorporate larger context without increasing the number of parameters. Dai *et al.* [7] exploit shape information via masking convolutional features. Hyeonwoo *et al.* [25] propose Deconvolution Network for Semantic Segmentation to identify detailed structures and handles objects in multiple scales naturally.

For human parsing, Yamaguchi *et al.* [37] tackle the clothing parsing problem using a retrieval based approach. Luo *et al.* [23] propose a Deep Decompositional Network for parsing pedestrian images into semantic regions. Liang *et al.* [16] propose a Contextualized Convolutional Neural Network to tackle the problem and achieve very impressing results. Xia *et al.* [35] propose the "Auto-Zoom Net" for human paring. Some other works explore how to jointly object and part segmentation using deep learned potentials [32]. Although great success achieved, these methods can not be directly applied in our setting where only one labeled frame per training video is available.

Weakly/semi-supervised semantic segmentation: Chen *et al.* [26] develop Expectation-Maximization (EM) methods to solve the semantic image segmentation from either weakly annotated training data or a combination of few strongly labeled and many weakly labeled images. Dai *et al.* [6] propose a method called "Boxsup" which only requires easily obtained bounding box annotations. Xu *et al.* [36] propose a unified approach that incorporates image level tags, bounding boxes, and partial labels to produce a pixel-wise labeling. Liu *et al.* [17] address the problem of automatically parsing the fashion images with weak supervision from the user-generated color-category tags. Wei *et al.* extend the weakly supervised classification solution [34] and propose a simple to complex framework for weakly-supervised semantic segmentation [33]. These methods have achieved competitive accuracy in weakly/semi supervised semantic segmentation but are not designed for video parsing task.

Optical flow v.s. semantic segmentation: Sevilla-Lara *et al.* [30] segment a scene into things, planes, and stuff and then pose the flow estimation problem using localized layers. Bai *et al.* [2] estimate the traffic participants using instance-level segmentation. The epipolar constraints is then used on each participant to govern each independent motion. In these methods, optical flow estimation benefits from semantic segmentation. However, SVP utilizes optical flow for better video parsing.

Pfister *et al.* [27] investigate a video pose estimation architecture that is able to benefit from temporal context by combining information across the multiple frames using optical flow. The key differences is that the optical flow is estimated offlined using dense optical flow while SVP is an end-to-end framework.

3. Approach

3.1. Framework

Suppose that we have a video $\mathcal{V} = \{I_1, \cdots, I_N\}$, where N is the number of frames. The single labeled frame is I_t, and its corresponding groundtruth is G_t. The pixel j of the labelmap P_t is denoted as P_t^j and takes the value within the range $[1, K]$, where K is the number of labels, such as "face", "bag" and "background".

The SVP network is shown in Figure 2. The input is a triplet $\{I_{t-l}, I_{t-s}, I_t\}$, among which only I_t is labeled. l and s are set empirically. The output is the parsing result P_t. SVP contains three sub-networks. As a preprocessing step, we use Faster R-CNN [28] to extract the human region. Then, the triplet are fed into Conv1$\sim$Conv5

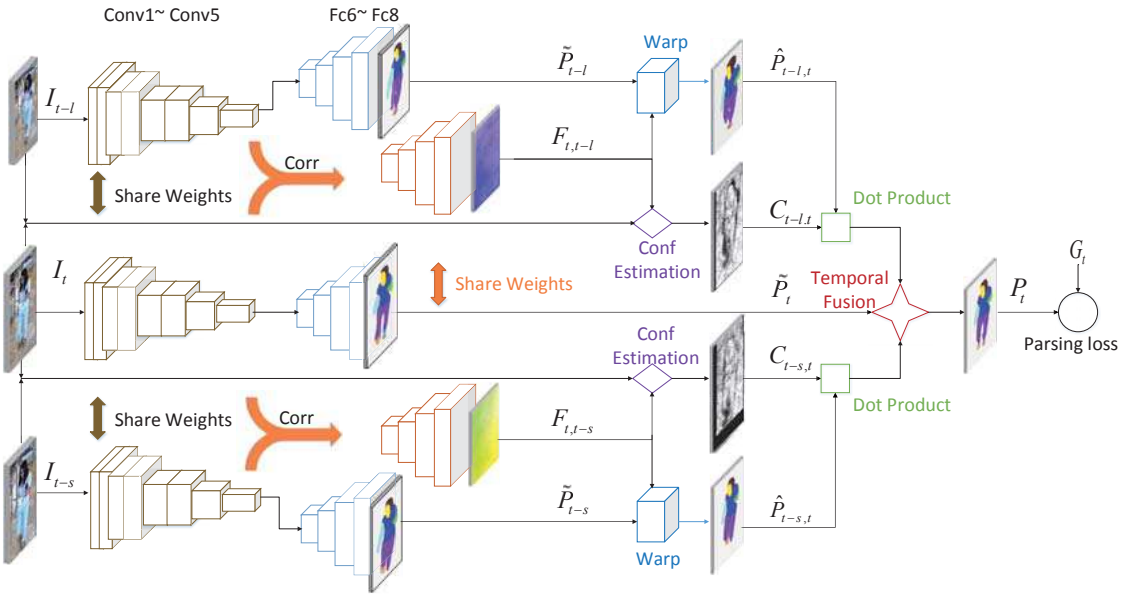

Figure 2. The proposed single frame supervised video paring (SVP) network. The network is trained end-to-end.

for discriminative feature extraction. The frame parsing sub-network (Section 3.2) produces the rough labelmaps for the triplet, denoted as $\{\tilde{P}_{t-l}, \tilde{P}_{t-s}, \tilde{P}_t\}$. The optical flow estimation sub-network aims to estimate the dense correspondence between adjacent frames (Section 3.3). The temporal fusion sub-network (Section 3.4) applies the obtained optical flow $F_{t,t-l}$ and $F_{t,t-s}$ to $\tilde{P}_{t-l}$ and $\tilde{P}_{t-s}$, producing $\hat{P}_{t-l,t}$ and $\hat{P}_{t-s,t}$. To alleviate the influence of imperfect optical flow, the pixel-wise flow confidences $C_{t-l,t}$ and $C_{t-s,t}$ are estimated. The quintet including $\left\{\tilde{P}_t, \hat{P}_{t-l,t}, \hat{P}_{t-s,t}, \mathrm{C_{t-l,t}}, \mathrm{C_{t-s,t}}\right\}$ are fused to produce the final P_t, upon which the softmax loss are defined. Extra supervision is also applied on $\hat{P}_{t-l,t}$ and $\hat{P}_{t-s,t}$ for better performance.

The image parsing and optical flow estimation sub-networks share first several convolution layers because the two tasks are implicitly correlated. More specifically, only pixels with the same labels can be matched by optical flow. Besides, both sub-networks make per pixel predictions. Frame parsing classifies each pixel while optical flow is the offset/shift of each pixel. Therefore, the optimal receptive fields of the two tasks are similar, which provides a prerequisite for feature sharing. The other benefit is to save a lot of computation.

3.2. Frame Parsing Sub-network

As shown in Figure 2, the frame parsing sub-network has three duplications with shared weights to deal with I_{t-l}, I_{t-s} and I_t respectively. The input is the 3-channel RGB image, and the output is the K channel confidence maps of the same resolution. In our experiments, DeepLab [4] is used. Our SVP framework is quite generic and is not limited to any specific image parsing method, other semantic segmentation methods [7, 21, 21, 1, 22] can also be used.

3.3. Optical Flow Estimation Sub-network

We resort to optical flow $F_{a,b} : R^2 \rightarrow R^2$ to build the pixel-wise correspondences between frames. The flow field $F^p_{a,b} = (q_x - p_x, q_y - p_y)$ computes the relative offsets from each point p in image I_a to a corresponding point q in image I_b. The optical flow estimation sub-network estimates the flow $F_{t,t-l} = o\,(I_t, I_{t-l})$, where $o\,(a, b)$ is the operation of predicting the optical flow from a to b. $F_{t,t-s}$ is estimated similarly. One feasible approach is to off-line calculate the optical flow via the state-of-the-art methods [3, 3] and load them into the network during optimization. It makes training and testing be a multi-stage pipeline, and thus very expensive in space and time. However, SVP computes the optical flow on the fly.

Network architecture: After the shared Conv1∼Conv5 layers, a "correlation layer" [10, 24] (denoted as "Corr" in Figure 2) performs multiplicative patch comparisons between two feature maps. After that, several "upconvolutional" layers are introduced to obtain the optical flow with the same resolution as the input image pairs. Since our surveillance dataset has no groundtruth optical flow, we use flying chairs dataset for training.

3.4. Temporal Fusion Sub-network

Optical flow confidence estimation: The optical flow estimated via the above mentioned method is imperfect. To

suppress noisy $\hat{P}_{t-l}$, we estimate the confidence of the estimated optical flow $F_{t,t-l}$ of each pixel. The flow $F_{t,t-s}$ can be handled in similar manners.

The flow confidence is defined based on the appearance reconstruction criterion [3]. Mathematically, for each pixel i in the optical flow $F_{t,t-l}$, its confidence $C^i_{t-l,t}$ is:

$$C^i_{t-l,t} = \left\| I^i_t - \hat{I}^i_t \right\|_1 = \left\| I^i_t - w^i \left(I_{t-l}, F_{t,t-l} \right) \right\|_1, \quad (1)$$

$\|\cdot\|_1$ denoted the L_1 norm. $\hat{I}^i_t$ is the wrapped counterpart of I^i_t. $w\left(a,b\right)$ is the operation of applying the estimated optical flow b to warp image a. The coordinates of pixel i in I_t is $\left(x^i, y^i\right)$, while the mapped coordinates in I_{t-l} is $\left(x^{i'}, y^{i'}\right) = \left(x^i, y^i\right) + F^i_{t,t-l}$. When $\left(x^{i'}, y^{i'}\right)$ falls into sub-pixel coordinate, we rewrite the $\hat{I}^i_t$ of Equation 1 via bilinear interpolation:

$$\begin{aligned} &\hat{I}^i_t = w^i \left(I_{t-l}, F_{t,t-l} \right) \\ &= \sum_{q \in \{neighbors\ of\ (x^{i'}, y^{i'})\}} I^q_{t-l} (1 - \left| x^{i'} - x^q \right|)(1 - \left| y^{i'} - y^q \right|), \end{aligned} \quad (2)$$

where q denotes the 4-pixel neighbors (top-left, top-right, bottom-left, bottom-right) of $\left(x^i, y^i\right)$.

The confidence defined in Equation 1 is the distance between the orignal image and its warped counterpart. The similaritiy is calcualted via:

$$C^i_{t-l,t} = exp(-C^i_{t-l,t}/2\sigma^2), \quad (3)$$

where σ is the mean value of $C_{t-l,t}$. Higher value indicates more confident optical flow estimation.

Temporal fusion: As shown in Figure 2, the estimated parsing results $\tilde{P}_{t-l}$ and $\tilde{P}_{t-s}$ are warped according to the optical flow $F_{t,t-l}$ and $F_{t,t-s}$ via:

$$\begin{aligned} \hat{P}_{t-l,t} &= w(\tilde{P}_{t-l}, F_{t,t-l}), \\ \hat{P}_{t-s,t} &= w(\tilde{P}_{t-s}, F_{t,t-s}). \end{aligned} \quad (4)$$

They are further weighted by the confidence map of Equation 1 to reduce the influence of inaccurate optical flow by: $\hat{P}_{t-l,t} \cdot C_{t-l,t}$ and $\hat{P}_{t-s,t} \cdot C_{t-s,t}$, where $\cdot$ denotes dot product. They are fused with $\tilde{P}_t$ via a temporal fusion layer with several 1×1 filters to produce the final P_t. To enforce accurate model training, we add extra/deep [14] supervision to $\hat{P}_{t-l,t}$, $\hat{P}_{t-s,t}$ and P_t.

3.5. Training Strategies

Like the Faster R-CNN [28], we adopt a 4-step alternating training algorithm for optimization. **(i)** we train the optical flow sub-network via the strategies in Section 3.3 with flying chairs dataset [10]. **(ii)** we train the frame parsing sub-network and the temporal fusion sub-network together using the optical flow estimated in step (i). Both optical flow and frame parsing sub-networks are initialized with VGG model [31]. The temporal fusion sub-network is initialized via standard Gaussian distribution (with zero mean and unit variance). At this point the two networks do not share convolutional layers. **(iii)** We fix the Conv1$\sim$Conv5 layers of optical flow estimation sub-network by those of frames parsing sub-network, and only fine-tune the layers unique to optical flow. Now the two sub-networks share convolutional layers. **(iv)** keeping Conv1$\sim$Conv5 layers fixed, we fine-tune the unique layers of frame parsing and temporal fusion sub-networks. As such, all sub-networks form a unified network.

The major reason of training the optical flow sub-network at the beginning is that, the temporal fusion sub-network *depends on* the optical flow results. Then, we replace the conv. layers of the optical flow sub-network by that of the parsing sub-network for three reasons. *First*, the two tasks are essentially correlated: parsing results are pixel-wise labels while optical flow is pixel-wise offset. Thus, the optimized conv. layers trained from the parsing network is expected to perform equally well for the optical flow network. *Second*, the optical flow sub-network is trained by an auxiliary flying chairs dataset, instead of the surveillance videos. Therefore, the conv. layers of the optical flow sub-network is less discriminative on our surveillance datasets. *Third*, the convolution layers need to be shared. Actually, we have attempted to train the whole network in a single phase, but found it hard to converge. We leave it as an important future work.

3.6. Inference

During inference, we slide a parsing window along the video to specifically consider the temporal context. The parsing results of I_t is jointly determined by the short video segment preceding it. For calculation simplicity, a triplet of frames, including long-range frame I_{t-l}, short-range frame I_{t-s} as well as I_t collaboratively contribute to the final parsing results P_t. Note that because the first l frames of a video do not have enough preceding frames to form a sliding parsing window, we apply the frame parsing sub-network alone to I_t and produce its parsing results.

4. Experiments

4.1. Experimental setting

Dataset & Evalutaion Metrics: Since there is no publicly available surveillance video parsing dataset, we manually build two datasets, one for indoor, the other for outdoor. The **indoor** dataset contains 700 videos, among which 400 videos and 300 videos are used as training set and test set, respectively. The **outdoor** dataset contains 198 training videos, and 109 testing videos. For both datasets, we randomly select and pixel-wisely label 1 frame from each training video. For each testing video, we randomly label 5

Table 1. Per-Class Comparison of F-1 scores with state-of-the-arts and several architectural variants of our model in Indoor dataset. (%).

Methods	bk	face	hair	U-clothes	L-arm	R-arm	pants	L-leg	R-leg	Dress	L-shoe	R-shoe	bag
PaperDoll [37]	92.62	57.16	58.22	62.52	19.96	14.99	52.47	25.43	20.7	9.92	20.66	24.41	14.32
ATR [15]	93.62	59.08	60.79	81.36	32.54	28.65	75.40	29.19	29.60	70.22	11.68	17.75	48.97
M-CNN [18]	93.40	53.94	59.12	75.53	24.46	20.51	78.46	36.15	21.92	43.61	14.53	18.79	53.43
Co-CNN [16]	94.06	64.64	73.53	81.54	26.82	31.66	77.13	25.47	34.11	76.08	15.42	20.57	46.91
FCN-8s [22]	94.80	71.35	74.90	79.53	33.55	32.29	81.89	36.57	33.98	43.53	33.03	31.50	43.66
DeepLab [4]	93.64	63.01	69.61	81.54	40.97	40.31	81.12	34.25	33.24	64.60	28.39	26.40	56.50
EM-Adapt [26]	93.46	66.54	70.54	77.72	42.95	42.20	82.19	39.42	37.19	63.22	33.18	31.68	53.00
SVP l	94.68	67.28	72.74	82.12	42.96	43.35	81.91	39.26	38.31	67.17	31.47	30.38	58.99
SVP s	94.65	66.27	73.48	83.12	45.17	44.89	82.72	38.62	38.43	66.04	30.93	31.46	58.81
SVP l+c	94.44	67.29	73.76	83.06	43.56	43.56	82.33	41.36	39.46	68.36	31.75	31.73	59.04
SVP s+c	94.64	67.62	74.13	83.48	45.13	**45.08**	83.21	39.89	40.11	68.17	31.15	32.27	58.75
SVP l+s	94.50	67.08	73.52	83.10	**45.51**	44.26	82.59	41.82	**42.31**	69.43	33.71	33.36	58.58
SVP l+s+c	**94.89**	**70.28**	**76.75**	**84.18**	44.79	43.29	**83.59**	**42.69**	40.30	**70.76**	**34.77**	**35.81**	**60.43**

Table 2. Per-Class Comparison of F-1 scores with state-of-the-arts and several architectural variants of our model in Outdoor dataset. (%).

Methods	bk	face	hair	U-clothes	L-arm	R-arm	pants	L-leg	R-leg	L-shoe	R-shoe	bag
FCN-8s [22]	92.00	62.64	65.58	78.64	28.73	28.97	79.69	38.88	9.08	32.04	30.56	29.45
DeepLab [4]	92.19	58.65	66.72	84.31	42.23	35.36	81.12	30.64	6.13	37.89	33.25	52.25
EM-Adapt [26]	92.68	60.84	67.17	84.78	41.28	33.61	**81.80**	**42.39**	7.28	39.54	32.20	**54.31**
SVP l	91.13	62.40	67.73	84.64	45.18	31.40	80.66	30.28	5.86	**40.32**	33.11	54.96
SVP s	92.51	64.25	67.14	84.99	45.28	32.14	79.71	32.31	18.49	37.24	31.45	51.58
SVP l+c	92.60	63.76	68.77	84.84	**45.83**	33.75	81.67	31.37	**19.06**	38.54	33.51	53.57
SVP s+c	**92.94**	64.40	69.93	**85.43**	44.44	31.86	81.65	35.88	18.22	37.48	33.36	54.23
SVP l+s	91.90	63.32	69.48	84.84	42.09	28.64	80.45	31.10	13.28	38.52	**35.52**	46.89
SVP l+s+c	92.27	**64.49**	**70.08**	85.38	39.94	**35.82**	80.83	30.39	13.14	37.95	34.54	50.38

frames for comprehensive testing. The indoor dataset contains 13 categories, namely face, hair, upper-clothes, left-arm, right-arm, pants, left-leg, right-leg, left-shoe, right-shoe, bag, dress, and background. The videos in the outdoor dataset are collected in winter, so the label "dress" is missing. To obtain human centric video, human are first detected via Faster R-CNN [28] fine-tuned on VOC dataset [9]. To speed up, we track the human by KCF[11]. Other tracking algorithms [39, 40, 20] can also be used. The obtained human centric images are fed into SVP.

We use the same metric as PaperDoll [37] to evaluate the performance. Among all evaluation metrics, the average F-1 is the most important metric. We train SVP via the Caffe [13] using Titan X. The initial learning rates for frame parsing and optical flow estimation sub-networks are 1e-8 and 1e-5 respectively. The long range l and short range s are empirically set as 3 and 1 in the indoor dataset. Because the outdoor dataset has a lower frame rate and contains more quick dynamics, l and s are set to 2 and 1.

4.2. Comparison with state-of-the-art

We compare our results with five state-of-the-art methods. The *1st* is **PaperDoll** [37]. It is the best traditional method. The *2nd* is **ATR** [15] formulating the human parsing task as an active template regression problem. The *3rd* baseline method is **M-CNN** [18], which is an quasi-Parametric human parsing method. The *4th* is **Co-CNN** [16] which uses a Contextualized Convolutional Neural Network to tackle the problem. The *5th* is **FCN-8s** [22], which achieves competitive results in several semantic segmentation benchmark datasets. The *6th* baseline is **DeepLab** [4]. The above mentioned three methods are supervised algorithms. Therefore, we only use the labeled set for training. The *7th* baseline method is **EM-Adapt**[1] [26] which can use both image-level and bounding-box annotation as weak- and semi-Supervised supervision. We also try another baseline DecoupledNet[2]. [12]. However, the results of DecoupledNet in both datasets are much lower than SVP and other baselines. The reason is that DecoupledNet first obtains the saliency map of each classified label. Deconvlution is then operated upon the map to generate the final parsing results. However, many labels, e.g., face, appear in almost every training image, which causes the classification network less sensitive to the position of these labels.

Table 3 shows the comparisons between SVP and 7 state-

[1] http://liangchiehchen.com/projects/DeepLab-LargeFOV-Semi-EM-Fixed.html

[2] http://cvlab.postech.ac.kr/research/decouplednet/

of-the-art methods in the **Indoor** dataset. Different variants of SVP are generated by gradually adding more components, which will be discussed in the next subsection. It can be seen that our best SVP, namely "SVP l+s+c" reaches the average F-1 score of 0.6020, which is superior than all baselines. The *1st* to the *6th* baselines all use labeled images. Therefore, the improvements show the advantage of utilizing the unlabeled dataset. EM-Adapt also uses unlabeled images, and thus reaches a higher F1-score of 0.5640, which is better than the six supervised baselines. However, EM-Adapt is still worse than all the variants of SVP. It shows that label propagation via optical flow is helpful in the surveillance video parsing task. The F1-scores of each category are shown in Table 1. We can observe that "SVP l+s+c" beats all baselines in all 13 categories, which again shows the big improvements brought by the proposed SVP.

Table 3. Comparison with state-of-the-arts and several architectural variants of our model in Indoor dataset. (%).

Methods	Accu	fg_accu	Avg.pre	Avg.rec	Avg. F-1
PaperDoll [37]	46.71	78.69	33.55	45.68	36.41
ATR [15]	85.69	71.24	47.39	53.21	49.14
M-CNN [18]	85.19	71.31	42.90	50.11	45.68
Co-CNN [16]	87.58	72.58	53.54	51.87	51.38
FCN-8s [22]	88.33	71.56	55.05	52.15	53.12
DeepLab [4]	86.88	77.45	49.88	**64.30**	54.89
EM-Adapt [26]	86.63	80.88	53.01	63.64	56.40
SVP l	88.81	74.42	56.28	59.81	57.74
SVP s	88.91	77.12	55.90	61.21	58.04
SVP l+c	88.75	77.28	56.07	61.94	58.43
SVP s+c	89.07	77.06	56.86	61.98	58.73
SVP l+s	88.85	**78.68**	56.77	62.73	59.21
SVP l+s+c	**89.88**	76.48	**61.52**	59.38	**60.20**

Among all the baselines, we find that FCN-8s, DeepLab and EM-Adapt show superior performances Therefore, we only compare with the 3 baselines in the **Outdoor** dataset. Table 4 shows the results. It can be seen that our method reaches the average F-1 score of 0.5294 while FCS-8s, DeepLab and EM-Adapt only reach 0.4433, 0.4775 and 0.4907. The improves are 0.0861, 0.0519 and 0.0387 respectively. Comparing Table 4 and Table 3, we find that the performances of all algorithms generally drop. The reason is that the outdoor dataset contains 198 training videos, while the number is doubled in the indoor dataset, reaching 400. The F1-scores of each category are shown in Table 2. We can observe that "SVP l+s+c" beats FCN and DeepLab in all 13 categories and is better than EM-Adapt in most categories, which again shows the effectiveness.

4.3. Component Analysis

Temporal fusion weights: We visualize the learned weights for the temporal fusion layers for R-arm and L-shoe in Figure 3 in the Indoor dataset. The horizontal axis has

Table 4. Comparison with state-of-the-arts and several architectural variants of our model in Outdoor dataset. (%).

Methods	Accu	fg_accu	Avg.pre	Avg.rec	Avg. F-1
FCN-8s [22]	82.46	70.70	43.22	50.09	44.33
DeepLab [4]	85.07	78.44	49.87	51.10	47.75
EM-Adapt [26]	85.82	76.87	50.82	52.98	49.07
SVP l	84.27	**81.51**	47.46	55.31	48.28
SVP s	85.83	73.48	53.46	50.63	49.01
SVP l+c	85.87	77.37	52.66	52.68	49.79
SVP s+c	**86.30**	77.13	52.89	**52.70**	49.99
SVP l+s	85.30	77.03	56.15	49.92	51.17
SVP l+s+c	85.71	79.26	**56.95**	52.14	**52.94**

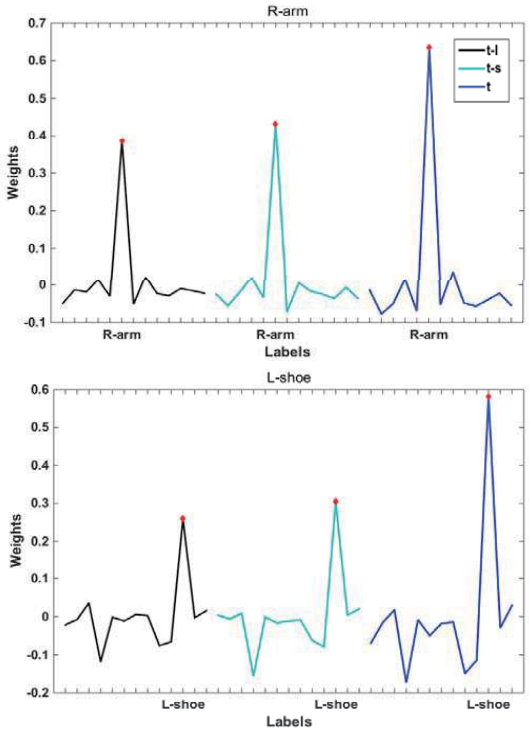

Figure 3. The temporal pooling weight for R-arm and L-shoe.

$3 \times K$ ticks, corresponding to the K labels for I_{t-l} (shown in black), I_{t-s} (shown in green) and I_t (shown in blue) sequentially. The vertical axis illustrates the fusion weights.

By analyzing the sub-figure for R-arm, we have several observations. First, the shapes of the weights for I_{t-l}, I_{t-s} and I_t are similar. Second, all maximum values for the triplet (denoted as red dots) are positive, which demonstrates that all frames contribute to the final result. Third, for all the frames, the labels reaching maximum values are all R-arm. Fourth, the maximum value of I_{t-s} is higher than that of I_{t-l}, because it contains less errors in optical flow. The maximum value of I_t is the highest, because it is the frame under consideration. Similar phenomenon can be found in the L-shoe case.

Long/Short range context: We test the effectiveness of long and short range frame. "SVP l" means SVP with long-

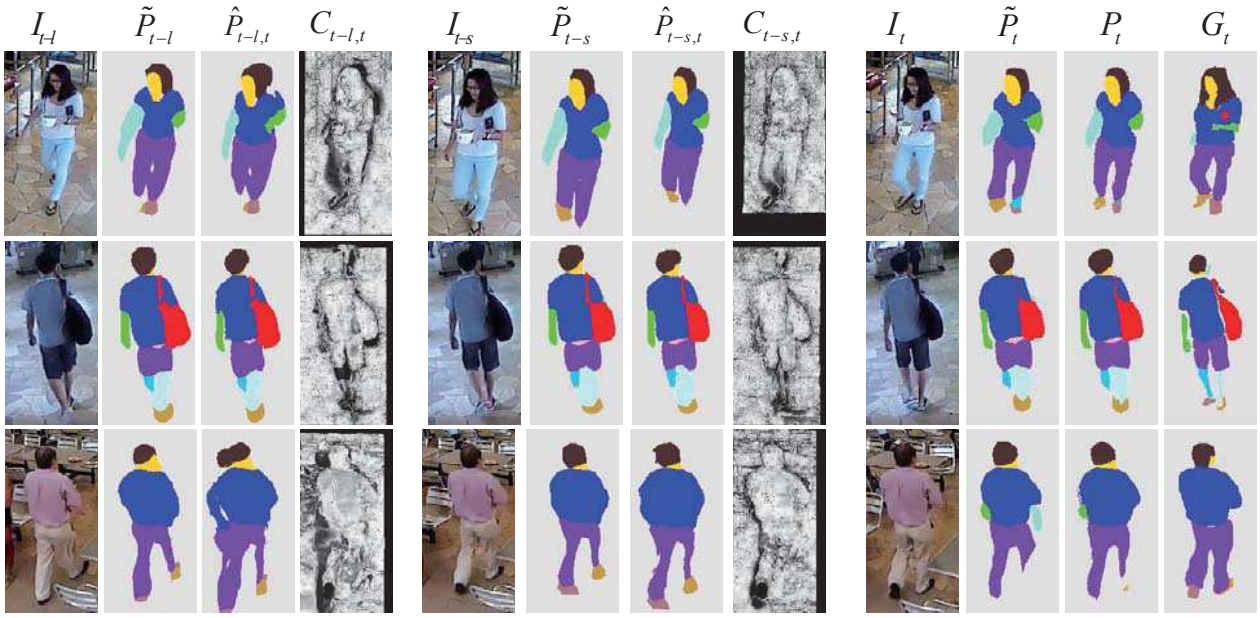

Figure 4. Step by step illustration of SVP. 1∼4 columns: the long-range frame, its the parsing result, the warped parsing result and the confidence map. 5∼8 columns: the short-range frame, its parsing result, the warped parsing result and the confidence map. 9∼12 columns: test image, the rough parsing result, refined parsing result and ground truth parsing result.

range context only. To implement this SVP variant, an image pair, namely I_t as well as I_{t-l} are fed into SVP during both training and testing phases. Similarly, "SVP s" is SVP containing only short-range frame. "SVP l+s" is the combination of them, meaning both long-range and short-range frames are considered. Table 3 shows the results in indoor dataset. The Ave.F-1 of "SVP l" and "SVP s" reach 0.5774 and 0.5804 respectively, which are lower than "SVP l+s" 0.5843. It proves the long and short range context are complementary. Similar conclusion can be drawn from outdoor dataset in Table 4. "SVP l" and "SVP s" achieve 0.4828 and 0.4901, while the combination of them reaches 0.4979. The per-class F1 score of "SVP l", "SVP s" and "SVP l+s" in indoor and outdoor datasets can be found in Table 1 and Table 2 respectively. They again show that both long and short range context are necessary.

Optical flow confidence: The flow confidence is designed for filtering/suppressing the noisy optical flow. To this end, we implement two SVP variants called "SVP l+c" and "SVP s+c" indicating either long or short-range optical flow is weighed by its confidence first and then contribute to the final parsing result. The results in indoor dataset is shown in Table 3. We find that "SVP l+c" improves "SVP l" and "SVP s+c" performs better than "SVP s". This demonstrates the effectiveness of optical flow confidence. The same conclusion can be drawn by comparing the F-1 score of "SVP l+s+c" and "SVP l+s". We also validate the effectiveness of optical flow confidence in outdoor dataset. As shown in Table 4, the F-1 score of "SVP l+s+c" is 0.5294, which is higher than "SVP l+s" 0.5117.

4.4. Qualitative Results

Figure 4 shows the stepwise results of SVP in indoor dataset. In the first row, the left shoe of the women is predicted as leg in $\tilde{P}_t$. The warped label from the I_{t-s}, denoted as $\hat{P}_{t-s,t}$ does not find left shoe. Thanks to the fusion from $\hat{P}_{t-l,t}$, the women's left shoe is labelled correctly in the final prediction P_t. Again in the first row, comparing with I_{t-s}, the women is far from the camera in I_t, and thus is relatively small. The foreground region shrinks from $\tilde{P}_{t-s}$ to $\hat{P}_{t-s,t}$, which shows that the estimated optical flow is very accurate. Inaccurate optical flow may result in the bad propagated parsing result, e.g., the shape of the hair in $\hat{P}_{t-l,t}$ is too large in the first row. However, the inaccurate hair region has a low confidence in $C_{t-l,t}$. Therefore, the fused result P_t has precise hair shape. In the second row, the strap of the bag is almost ignored in $\tilde{P}_t$. However, both $\tilde{P}_{t-l}$ and $\tilde{P}_{t-s}$ find the strap, and help to distinguish the strap from the upper-clothes successfully in P_t. In the third row, the P_t correctly removes the wrongly predicted arm in $\tilde{P}_t$. The I_{t-l} is not warped very good, and there is a ghost behind this man in the labelmap $\hat{P}_{t-l,t}$. But fortunately it does not affect the fused prediction P_t, because the confidence of this ghost is very low in $C_{t-l,t}$ and hence it is filtered out during the fusion.

Several qualitative results of both datasets are shown in Figure 5. The first three rows show paring results of the indoor dataset while the last two rows demonstrate those of outdoor dataset. In each group, the test image, the groundtruth, the parsing results of EM-Adapt and SVP are

Figure 5. The test image, the groundtruth label, results of the EM-Adapt and SVP are shown sequentially.

shown. It can be seen that SVP is generally better than EM-Adapt from two aspects. First, SVP correctly estimates the existence of a label. For example, for the image in the second row second column, the region wrongly predicted as upper clothes by EM-Adapt is correctly predicted as dress by SVP. Another example is second row first column. EM-Adapt misses the left shoe. SVP correctly predicts the left shoe's existence and location. Second, SVP can better estimate the shape of the labels. For example, in the first image in top row, the shape of the bag strap is slender, which is correctly estimated by SVP. Moreover, the shapes of shoes estimated by SVP are more accurate than EM-Adapt. For another example, SVP better identifies the shapes of pants and left/right arms in the third image of the third row.

4.5. Time Complexity

Note that in the inference stage, much computation can be saved. For example, when parsing frame I_t, the long-range frame I_{t-l} and short-range frame I_{t-s} do not need go through the frame parsing sub-network because their rough parsing results P_{t-l} and P_{t-s} have already been calculated. For another example, the extra computation brought by the optical flow estimation sub-network is small because the Conv1$\sim$Conv5 features are shared. Moreover, the fusion layer contains several 1×1 convolutions and thus is not quite time-consuming.

5. Conclusion & Future Works

In this work, we present an end-to-end single frame supervised video parsing network. To parse a testing frame, SVP processes a video segment preceding it. The rough frame parsing results and the on-line computed optical flows among frames are fused to produce refined parsing results. We demonstrate the effectiveness of SVP on two newly collected surveillance video parsing datasets.

In future, we will build an online demo to parse any surveillance video uploaded by users in real time. Moreover, we plan to apply SVP to parse other kinds of videos, such as urban scene videos [5].

Acknowledgment

This work was supported by National Natural Science Foundation of China (No.U1536203, Grant 61572493, 11301523), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) 201600035. We also would like to thank NVIDIA for GPU donation.

References

[1] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv:1505.07293*, 2015.

[2] M. Bai, W. Luo, K. Kundu, and R. Urtasun. Exploiting semantic information and deep matching for optical flow. In *ECCV*, 2016.

[3] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *TPAMI*, 2011.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv:1412.7062*, 2014.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. *arXiv:1604.01685*, 2016.

[6] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[7] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015.

[8] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM MM*, 2014.

[9] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015.

[10] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. *arXiv:1504.06852*, 2015.

[11] J. F. Henriques, C. Rui, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 2014.

[12] S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *NIPS*, 2015.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[14] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *AISTATS*, volume 2, page 6, 2015.

[15] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, L. Lin, and S. Yan. Deep human parsing with active template regression. *TPAMI*, 2015.

[16] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. *ICCV*, 2015.

[17] S. Liu, J. Feng, C. Domokos, and H. Xu. Fashion parsing with weak color-category labels. *TMM*, 2014.

[18] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-cnn meets knn: Quasi-parametric human parsing. *arXiv:1504.01220*, 2015.

[19] S. Liu, Z. Song, G. Liu, S. Yan, C. Xu, and H. Lu. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012.

[20] S. Liu, T. Zhang, X. Cao, and C. Xu. Structural correlation filter for robust visual tracking. In *CVPR*, 2016.

[21] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *ICCV*, 2015.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015.

[23] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In *ICCV*, 2013.

[24] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *CVPR*, 2016.

[25] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[26] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

[27] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. *arXiv:1506.02897*, 2015.

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[29] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *CVPR*, 2015.

[30] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black. Optical flow with semantic segmentation and localized layers. *arXiv:1603.03911*, 2016.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[32] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Joint object and part segmentation using deep learned potentials. In *ICCV*, pages 1573–1581, 2015.

[33] Y. Wei, X. Liang, Y. Chen, X. Shen, M. M. Cheng, J. Feng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI*, 2015.

[34] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Hcp: A flexible cnn framework for multi-label image classification. *TPAMI*, 2015.

[35] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human part segmentation with auto zoom net. *arXiv:1511.06881*, 2015.

[36] J. Xu, A. G. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *CVPR*, 2015.

[37] K. Yamaguchi, M. H. Kiapour, and T. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013.

[38] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *ACM MM*, 2013.

[39] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. *Low-Rank Sparse Learning for Robust Visual Tracking*. 2012.

[40] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via multi-task sparse learning. 2012.

[41] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014.